РОССИЙСКАЯ АКАДЕМИЯ НАУК

Проект

Создание вычислительной системы для моделирования суперкомпьютера с производительностью экзафлопсного уровня

Утверждено академиком-секретарем отделения нанотехнологий и информационных технологий РАН, академиком Велиховым Е.П.

и заместителем академика-секретаря отделения математических наук РАН, академиком Жижченко А.Б.

Руководитель проекта научный руководитель ФГУП «НИИ «Квант», академик Левин В.К.

Соруководитель проекта директор ФГУП «НИИ «Квант»
Елизаров Г.С.

Соруководитель проекта директор ИПМ им.М.В.Келдыша РАН, академик Четверушкин Б.Н.

Введение

Цель проекта — создание ВС для моделирования суперкомпьютеров со сверхвысоким уровнем распараллеливания вычислений в рамках работ по созданию в РФ вычислительных систем с производительностью на уровне 10^{18} оп./с.

Предпосылки реализации проекта

Создание суперкомпьютеров, имеющих производительность на уровне 10 экзафлопс (10^{18} операций с плавающей точкой в секунду — flops), к 2018-2020 годам признано в качестве стратегической научно-технической задачи в наиболее развитых странах мира — в США, Китае, Японии, странах ЕС, России, Индии.

Исходя из современных представлений о развитии элементной базы (сверхбольших интегральных схем и соответствующей памяти) и системных архитектур, суперкомпьютер экзафлопсного уровня производительности должен обрабатывать большое количество потоков данных (порядка 10^9 при производительности в одном потоке порядка 10^9 flops) при использовании общей памяти или секционированной памяти. Могут использоваться как асинхронные потоки, протекающие в универсальных процессорах и имеющие собственные счётчики команд, так и синхронные потоки, управляемые общим счётчиком команд для разных арифметико-логических устройств. Например, в графических процессорах-ускорителях Nvidia Fermi синхронно в каждом такте может протекать до 512 потоков. Синхронные потоки аппаратно более экономичны, но сложнее в программировании.

Модель суперкомпьютера

Объектом моделирования является суперкомпьютер экзафлопсного уровня производительности, представленный в виде узлов, соединенных одной или несколькими высокопроизводительными коммуникационными сетями.

Модели суперкомпьютера должны отвечать следующим требованиям:

- Сохранение всех ожидаемых архитектурных особенностей узлов суперкомпьютера экзафлопсного уровня производительности, таких

как количество процессоров, количество процессорных ядер, соотношение объема памяти и объема вычислений в узле, количество независимых вычислительных потоков.

- Настраиваемость архитектуры: гибкость, достаточная для быстрой модификации архитектуры процессорных ядер, типов и топологий внутри процессорных коммуникационных сетей, архитектуры межпроцессорных соединений и контроллеров памяти для моделирования различных перспективных суперкомпьютерных архитектур будущего в пределах, прогнозируемых для систем 2018 года.

Модель должна быть выполнена в некотором масштабе производительности, достаточном для создания рассматриваемого прототипа в 2013-14 годах. Для адекватного моделирования модель должна содержать соизмеримое с оригиналом количество ядер, не менее 1%, желательно 3-5% от количества ядер суперкомпьютера экзафлопсного уровня производительности.

Описание проекта

Создаваемая моделирующая система должна содержать наиболее важные технические и программные решения моделируемого суперкомпьютера. По сути, модель – это своего рода суперкомпьютер, воплощающий архитектурные идеи вычислительной системы экзафлопсного уровня производительности и позволяющий оценивать их эффективность, в том числе, путем составления, отладки и оценки производительности системного и прикладного ПО.

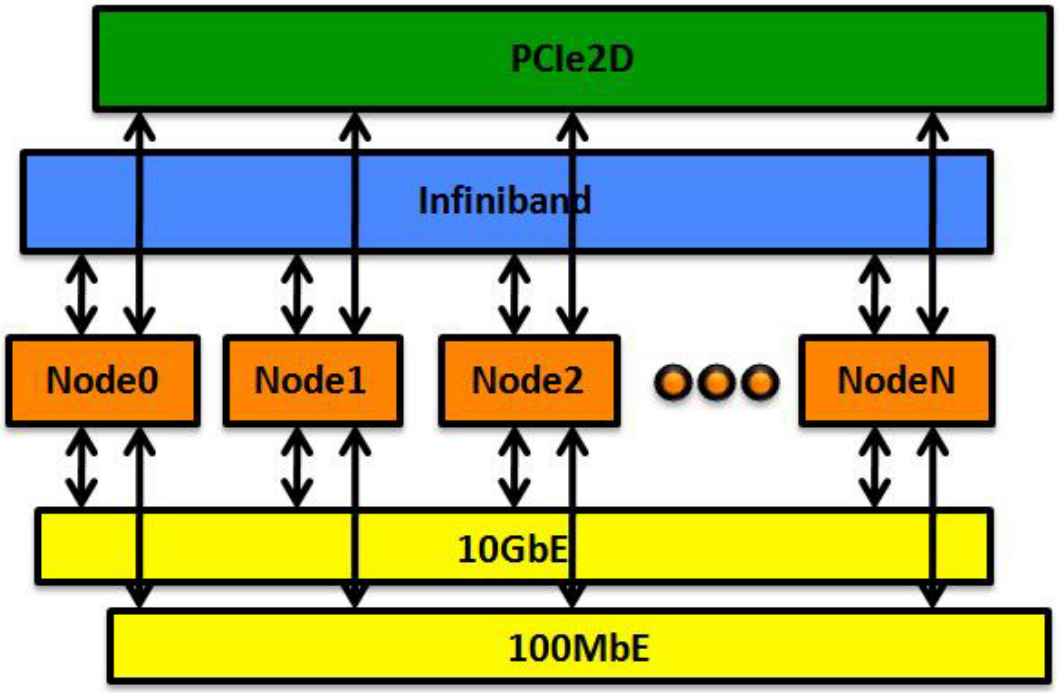

Рис. 1. Общая структура

Node0..N – вычислительные узлы (до 320 шт.)

Infiniband — сеть FDR InfiniBand, топология «толстое дерево», (сеть MPI-вычислений, сеть синхронизации, внешний сегмент сети хранения данных)

PCIe2D — 2D решетка до 16x10, каждая строка и столбец объединены звездой отдельными коммутаторами PCIe Gen3 4x (реализуют сеть PGAS, общая память)

10GbE — 1/10 GbEthernet сеть загрузки и управления

100MbE – 100Mb Ethernet сеть мониторинга

Встроенные в процессорные кристаллы контроллеры памяти должны поддерживать высокую степень расслоения локального блока памяти, а процессорные ядра допускать большое число незавершённых обращений к памяти. На уровне суперкомпьютера должно программно формироваться глобальное адресное пространство разделяемой памяти, состоящей из блоков локальной памяти узлов, имеющих достаточно большой объем.

В каждом ВУ к портам коммутатора PCI Express подключаются ускорители на ПЛИС. Ускоритель представляет собой одну или несколько ПЛИС, к которым подключены блоки памяти. Предусматривается, что в каждой ПЛИС реализуются контроллеры блоков памяти, подключенных к ней, а также контроллер PCI Express. Предполагается, что функциональность контроллера памяти ПЛИС должна быть расширена для обеспечения работы процессорных ядер «по готовности данных» на аппаратном уровне.

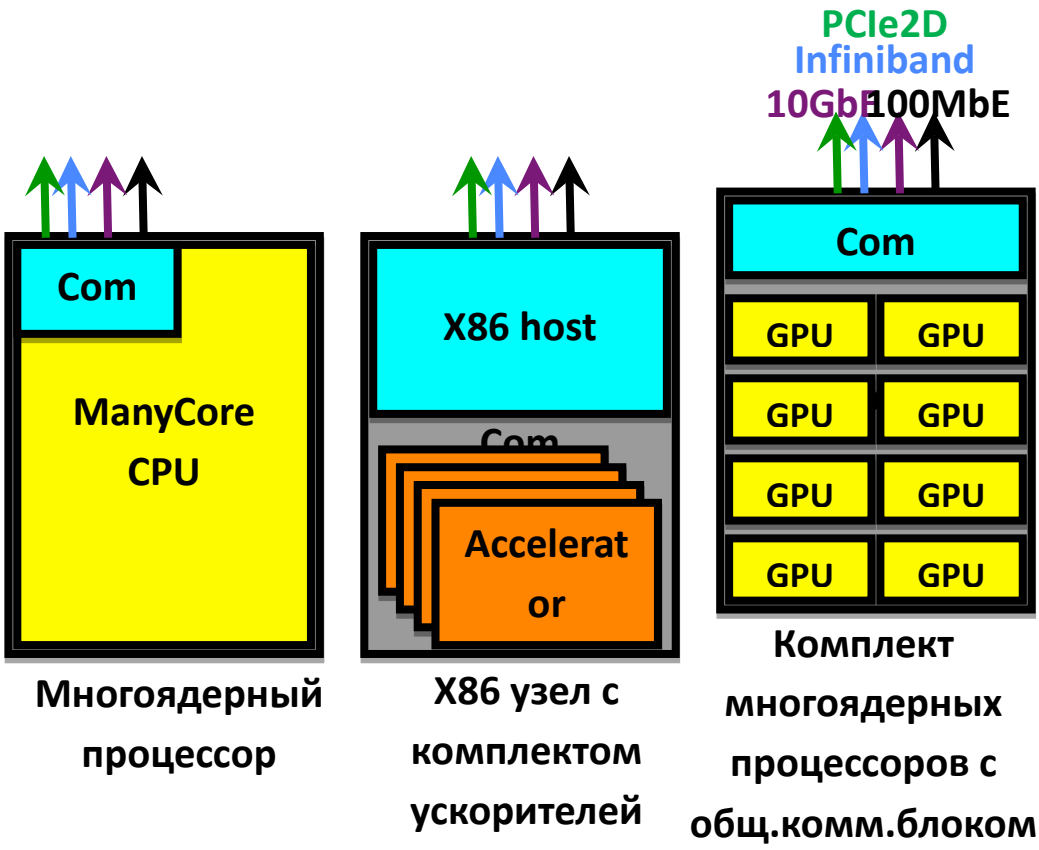

Рис. 2. Структура вычислительного узла

Модель программирования использует явное задание потоков, синхронизация которых реализуется с помощью дополнительных признаков данных в общей памяти и оригинального контроллера памяти ПЛИС. Для порождения и завершения асинхронных потоков в программах на традиционных языках программирования, например на языке Си, могут быть использованы дополнительные команды или библиотечные функции,

позволяющие соответственно породить заданное количество потоков, завершить потоки, а также выполнить атомарную операцию. В совокупности с общей памятью, слова которой снабжены дополнительными признаками и соответствующей функциональностью контроллера памяти, это позволяет задать синхронизацию и практически любые межпотоковые коммуникации.

На одном ПЛИС семейства Virtex7 будет реализован многоядерный многопотоковый процессор с общей памятью, содержащий порядка тысячи предельно компактных многопотоковых 32-х разрядных RISC-ядер, аппаратная часть которых позволяет порождать, манипулировать и завершать потоки, не обращаясь к системным вызовам ОС. Контроллер памяти будет поддерживать работу процессорных ядер «по готовности данных» на аппаратном уровне и механизмы межпотоковой синхронизации.

На указанный процессор будет портирована операционная система типа Minix и все стандартные инструменты для включения такого процессора в гетерогенную вычислительную систему.

Таким образом, модель, включающая 10000 ПЛИС, сможет достаточно всесторонне моделировать работу многопроцессорной вычислительной системы, содержащей несколько миллионов ядер, что соответствует представлению о будущей экзафлопсной ВС.

Сроки и этапы реализации проекта

Проект в целом:

3 кв. 2012 г. – 4 кв. 2016 г.

1 этап – разработка эскизно-технического проекта:

3 кв. 2012 г. – 4 кв. 2013 г.

2 этап – изготовление 1-ой очереди моделирующей гетерогенной ВС:

1 кв. 2014 г. – 2 кв. 2015 г.

3 этап – изготовление 2-ой очереди моделирующей гетерогенной ВС:

3кв. 2015 г. – 4кв. 2016 г.

Распределение работ

ИПМ им. Келдыша РАН – координация работ. Разработка принципов сетевого объединения вычислительных ресурсов. Разработка системного ПО. Разработка тестового ПО для моделирования вычислительных процессов экзафлопсных суперкомпьютеров. Проведение работ по моделированию экзафлопсных суперкомпьютеров.

ФГУП «НИИ «Квант» – разработка проекта. Разработка аппаратных средств. Системная интеграция и комплексная наладка.

Для разработки средств моделирования на отдельных ПЛИС фрагментов многоядерных кластерных вычислительных средств предполагается привлечь Центр инженерной физики МГУ имени М.В. Ломоносова.